LARGE VOCABULARY DECODING AND CONFIDENCE ESTIMATION USING WORD POSTERIOR PROBABILITIES

G. Evermann & P.C. Woodland

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK
Email: {ge204,pcw}@eng.cam.ac.uk

ABSTRACT

This paper investigates the estimation of word posterior probabilities based on word lattices and presents applications of these posteriors in a large vocabulary speech recognition system. A novel approach to integrating these word posterior probability distributions into a conventional Viterbi decoder is presented. The problem of the robust estimation of confidence scores from word posteriors is examined and a method based on decision trees is suggested. The effectiveness of these techniques is demonstrated on the broadcast news and the conversational telephone speech corpora where improvements both in terms of word error rate and normalised cross entropy were achieved compared to the baseline HTK evaluation systems.

1. INTRODUCTION

The output of a speech recognition decoder is typically a word lattice which contains a large number of competing word hypotheses and their associated likelihoods. These likelihoods are used to rank competing hypotheses and to select the 1-best output of the recogniser. In a Viterbi decoder only the best scoring (state level) path is considered and the likelihoods of all other paths (i.e. different time segmentations of the same word sequence or competing word hypotheses) have no direct influence on the decoder decision. This paper investigates techniques to augment these likelihood with estimates of word level posterior probabilities that allow information about these alternative paths to be incorporated into the decoder.

The following section discusses the estimation of word level posterior probabilities from word lattices generated by a Viterbi decoder. In section 3, a modified decoding approach is presented that incorporates the word posterior information into the search. The application of word posteriors to the problem of confidence scoring is investigated in section 4 and in the final section experimental results of an implementation of these techniques based on the HTK broadcast news (Hub4) and conversational telephone speech (Hub5) evaluation systems are presented.

2. ESTIMATING WORD LEVEL POSTERIOR PROBABILITIES

The estimation of word level posterior probabilities is based on the scores contained in a word lattice. Each link in the lattice is labelled with the acoustic and language mode likelihoods together with the start and end times of a particular word hypothesis. Due to the context dependency of the acoustic and language models it is necessary to ensure that for each link the relevant context is unique. This results in a large number of links labelled with the same word but corresponding to different segmentations and contexts.

The word posterior estimation is performed in two steps. First the well known forward-backward algorithm is used to calculate a link posterior probability for each link in the lattice. The link posterior $p(l|\boldsymbol{X})$ is defined as the sum of the probabilities of all paths $\boldsymbol{q}$ passing through the link l normalised by the probability of the signal $p(\boldsymbol{X})$:

$$p(l|\boldsymbol{X}) = \frac{\sum_{\boldsymbol{Q}_l} p(\boldsymbol{q},\boldsymbol{X})}{p(\boldsymbol{X})} \tag{1}$$

where $p(\boldsymbol{X})$ is approximated by the sum over all paths through the lattice. The probability of a path $p(\boldsymbol{q},\boldsymbol{X})$ is composed from the acoustic likelihood $p_{acc}(\boldsymbol{X}|\boldsymbol{q})$ and the language model likelihood $p_{lm}(\boldsymbol{W})$:

$$p(\boldsymbol{q},\boldsymbol{X}) = p_{acc}(\boldsymbol{X}|\boldsymbol{q})^{\frac{1}{\gamma}} p_{lm}(\boldsymbol{W}) \tag{2}$$

Here it is important not to scale up the language model likelihoods as is usually done in a Viterbi decoder but instead to scale down the acoustic likelihoods, as the acoustic model severely underestimates the emission probabilities due to invalid independence assumptions. In all the experiments reported here the acoustic model scale factor was taken as the reciprocal of the standard language model factor.

One word in the speech signal is usually represented by a large number of links in the lattice (corresponding to different segmentations or language model contexts). When the segmentation or the surrounding context is not of interest it is necessary to combine the posteriors of these links to get a reliable estimate of the actual *word* posterior probability.

Such a word posterior can be seen as the generalisation of the N-best list based *log-likelihood ratio* introduced in [7] to word lattices. In the lattice based case, the same problem discussed in [7] is encountered, namely to determine which occurrences of the same word in different paths (or N-best list entries respectively) to consider as belonging to the same word instance in the speech signal.

To appear in *Proc. ICASSP2000, June 05-09, 2000, Istanbul, Turkey*

A possible solution is to define a word posterior distribution for each time frame by adding the link posteriors of all links spanning a given frame which correspond to the same word (this is effectively the same approach as the one suggested in [8]). This distribution is well defined as the link posteriors of all links spanning a frame sum to one. These *time dependent word posteriors* can be used either directly or can be combined to get one word posterior for each link (e.g. [8] suggests picking the maximum value seen in the time interval covered by the link). Here, the geometric mean of the time dependent posteriors in the time interval of the link is used.

An alternative solution is to use the clustering procedure proposed in [5] in the framework of "consensual lattice post-processing". In that case the posteriors of time overlapping links corresponding to the same word are added to yield word posterior estimates.

3. DECODING

The estimates of the word posteriors can be incorporated into the decoding process to improve the recognition accuracy. The advantage of posterior scores over conventional (Viterbi) likelihoods is that they consider not only the best segmentation of the hypothesis but also incorporate information about other segmentations of the same hypothesis and the relative likelihoods of all competing alternatives.

3.1. Posterior Rescoring

By examining the time dependent posterior distributions it was found that a reasonable decoder could be implemented by just picking the best word from the distribution in each frame. The problem with this approach is that it would completely disregard word sequence constraints by treating all frames independently and would, for example, not be able to detect when the same word is spoken twice in sequence.

To avoid these problems, the word posterior distributions were combined with the conventional Viterbi scores contained in the lattice. Inspired by the use of confidence scores for rescoring in [2], the word posteriors were added as an additional score to the acoustic and language model scores and the normal A* search was performed based on the resulting new decoder objective function:

$$f(\boldsymbol{W}) = p(\boldsymbol{X}, \hat{\boldsymbol{q}}|\boldsymbol{W})^{\frac{1}{\gamma}} p(\boldsymbol{W}) \rho^{\frac{\|\boldsymbol{W}\|}{\gamma}} \prod_{t=0}^{T} p(w(\hat{\boldsymbol{q}}, t), t|\boldsymbol{X}) \quad (3)$$

where γ is the language model weight, ρ the word insertion penalty and $\|\boldsymbol{W}\|$ the number of words in the hypothesis. Here, $p(w(\hat{\boldsymbol{q}}, t), t|\boldsymbol{X})$ is the word posterior probability of the word hypothesised at time t in path $\hat{\boldsymbol{q}}$.

Thus the word posteriors act as a local consistency measure, if one link hypothesis is supported by many high scoring alternatives then its likelihood is increased.

3.2. Minimum Word Error Rate Decoding

As an alternative to the posterior rescoring described above, the lattice post-processing approach presented in [5] was also investigated. This technique aims to compensate for the mismatch between the sentence based decoder objective function (maximum a-posteriori) and the word based evaluation metric. It has been suggested in [6] that while the MAP criterion leads to a decoder that finds hypotheses which are optimal in terms of the sentence error rate (SER) this is not necessarily optimal with respect to minimising the word error rate (WER).

The idea behind the algorithm in [5] is to first merge links belonging to the same word in the same time segment in order to obtain a word posterior estimate and then transform the resulting lattice into a linear graph (called *confusion network*) in which all paths pass through all nodes. This transformation is performed by a clustering procedure that groups time overlapping links into clusters based on their phonetic similarity while preserving the precedence order of the links encoded in the original lattice. This procedure is repeated until a total order of the links is achieved (i.e. two links are either in the same cluster or one precedes the other). By picking the hypothesis with the highest posterior probability from each cluster, the word sequence that minimises the expected word error rate can be found (according to the posterior distribution of word sequences represented by the lattice).

3.3. Experiments

Preliminary experiments were conducted on the lattices generated by the HTK system used in the 1997 Hub4 (broadcast news) and the 1998 Hub5 (conversational telephone speech) evaluations using triphone acoustic models and 4-gram language models (see the system descriptions in [10] and [4] respectively for details).

	Hub4		Hub5	
	WER	SER	WER	SER
baseline	17.4	92.0	42.6	80.2
post	17.0	92.0	42.5	80.5
confnet	16.9	92.3	41.5	80.6

Table 1: Decoding experiments on triphone lattices for HTK Hub4/Hub5 systems using time dependent posteriors (post) and confusion network clustering (confnet)

These experiments clearly show that word based posteriors can be used to improve the accuracy of a Viterbi MAP decoder. The results also exhibit the expected tradeoff between word and sentence error rates, i.e. a decrease in WER but an increase in SER. The confusion network technique proved to be more more robust and yields similar improvements on both corpora while the posterior rescoring technique worked well on the Hub4 corpus but gave no significant improvement on the Hub5 data. Obviously it is more important to allow a flexible time alignment of hypotheses on the Hub5 data where the time segmentation tends too be rather poor. While the posterior rescoring only considers links covering the same frame, the confusion network clustering explicitly "moves" hypotheses in time to find the optimal alignment based on the phonetic similarity of competing hypotheses thus compensating for the poor segmentation performed by the acoustic models.

In contrast to the experiments described in [6] we found the explicit WER minimisation using word level posteriors to be effective for the broadcast news system despite its relatively high overall accuracy. It is also interesting to note that the improvement is consistent over the various types of data found in broadcast news.

The reasoning in [6] is that at lower word error rates the correlation between sentence and word error rates should be much stronger than at higher error levels. In the segmentation used in the HTK Hub4 system the average sentence length is 38 words while it is only 9 words for the Hub5 corpus.[1] Our baseline system gave an average of 8.4 word errors per sentence on Hub4 while the average on Hub5 is only 4.1 words. The higher number of word errors per sentence certainly leads to a weaker correlation between the SER (which is optimised by a MAP decoder) and the WER (optimised by a word posterior based decoder) and therefore offers a greater potential for improvements by using a word based metric in the decoder even at the lower word error rate level (see [1] for a more detailed discussion).

4. CONFIDENCE SCORES

For many application it is very useful to annotate the 1-best result found by the decoder with confidence scores indicating how certain the system is about each word hypothesis. The standard metric used to asses the quality of a set of such confidence tags is the normalised cross entropy (NCE) which is an information theoretic measure of how much additional information the tags provide over the trivial baseline case of tagging all words with the same (optimal) score. This metric is used by NIST in the official scoring of evaluations and will be quoted for all the following experiments.

As proposed in [9], word posterior probabilities can be used directly as confidence scores of the word hypotheses. In our experiments we found that the lattice based methods tend to overestimate the posterior probabilities of words. The posteriors are therefore relatively poor confidence scores especially if the lattices used are small and contain only a small fraction of the likely word sequences.

In Table 2 the quality of the posterior probabilities used directly as confidence scores is compared across different corpora and systems. The lattices used were generated by the HTK systems mentioned in section 3. All experiments were run separately on the outputs of the triphone and quinphone stages (P3 and P7 respectively in [4]).

	triphone	quinphone
Hub4 eval'97	0.302	0.163
Hub5 dev'98	0.191	-0.026
Hub5 eval'98	0.104	-0.200

Table 2: Normalised cross entropy of the word posteriors based on time dependent posteriors

It can be seen that for the quinphone systems, with relatively smaller lattices, the posterior probabilities are less useful as confidence scores than on the triphone systems. This effect is significantly more pronounced on the Hub5 data because here the acoustic models are not able to distinguish as sharply between the highest scoring hypothesis and the competing alternatives as the Hub4 models, resulting in a larger number of word sequences with similar high likelihoods. Therefore the limited lattice size has a greater impact than for the Hub4 system as a larger proportion of the relevant (high scoring) word sequences are pruned.

The influence of the lattice size was further investigated by pruning the triphone lattices at a number of pruning thresholds and calculating word posteriors based on these pruned lattices. It was found that below a certain lattice size the NCE rapidly deteriorated as the average posterior value increased with decreasing lattice size. In smaller lattices there are often time segments where all paths pass through links corresponding to the same word resulting in a posterior estimate of 1.0 for this word.[2]

To compensate for the effects of the lattice size and the resulting overestimation of the posteriors a decision tree was trained for each system to map the posterior probabilities to confidence scores. Based on the step function defined by the decision tree a piecewise linear mapping function was chosen and applied to the posterior values. Table 3 summarises the performance of the resulting confidence scores for both methods of posterior estimation described in section 2. Quite clearly the mapping is necessary for both methods especially for the quinphone based system with its smaller lattices. The NCE on the quinphone systems is still significantly lower than on the triphone system because obviously the information contained in the part of the posterior distribution that has been pruned cannot be recovered by the tree based mapping.

	triphone		quinphone	
	post	tree	post	tree
time dep.	0.104	0.234	-0.200	0.188
confnet	0.000	0.213	-0.396	0.198

Table 3: NCE of confidences scores on Hub5 '98 evaluation system based on time dependent posteriors and confusion network clustering with and without tree mapping

5. FULL RECOGNITION RESULTS

The following experiments asses the effect of the techniques described above when implementing them in the full HTK evaluation systems used in the 1997 Hub4 and the 1998 Hub5 evaluations. Both systems operate in multiple stages, employ word 4-gram language models[3] and triphone and quinphone acoustic models adapted using MLLR. A full description of the stages involved can be found in [10] and [4] respectively.

The lattices generated by the best triphone and quinphone systems were rescored using the techniques discussed in section 3. The results for the Hub4 system are given in table 4. While both techniques are similarly successful

[1] Here "sentence" refers to the segments of the speech signal that the decoder works on, not necessarily linguistically meaningful units.

[2] Although the confidence scores are usually limited to some maximum value (e.g. $1.0 - 10^{-7}$) assigning this value to an incorrectly recognised word has a devastating impact on the NCE.

[3] The Hub5 system also used a class based trigram model.

in achieving a useful improvement on the triphone lattices, their effectiveness on the quinphone system is much smaller.

	triphone	quinphone
baseline	17.4	16.2
post-dec	17.0	16.1
confnet	16.9	16.0

Table 4: WER for Hub4 eval'97 decoding experiments

For the Hub5 system, in addition to the rescoring experiments we performed the improved confidence score estimation described in section 4 and generated the final system output by combining the best triphone based system with the best quinphone systems using Rover (see [3]). The baseline results given in Table 5 were produced by extracting the highest scoring word sequence from the lattices using the conventional Viterbi MAP criterion and generating confidence scores based on the N-best homogeneity measure.

	triphone	quinphone	Rover
baseline	42.6 (0.182)	40.3 (0.170)	39.5 (0.145)
post-dec	42.5 (0.234)	40.0 (0.188)	39.1 (0.197)
confnet	41.5 (0.213)	39.7 (0.198)	39.1 (0.186)

Table 5: WER & NCE for the Hub5 eval'98 set

Although the posterior rescoring approach (post-dec) performs more poorly than the confusion network decoder in terms of word error rate on the individual systems, the better confidence estimates seem to compensate for this and both techniques achieve the same WER improvement over the baseline on the combined system.

Both word posterior based approaches to confidence score estimation outperform the N-Best homogeneity measure in all experiments. As an example, Figure 1 shows a comparison of the DET curves for the confidence scores in the baseline system (after combination with Rover) and the time dependent posterior based one. At all operating points the time dependent posterior technique gives better performance.

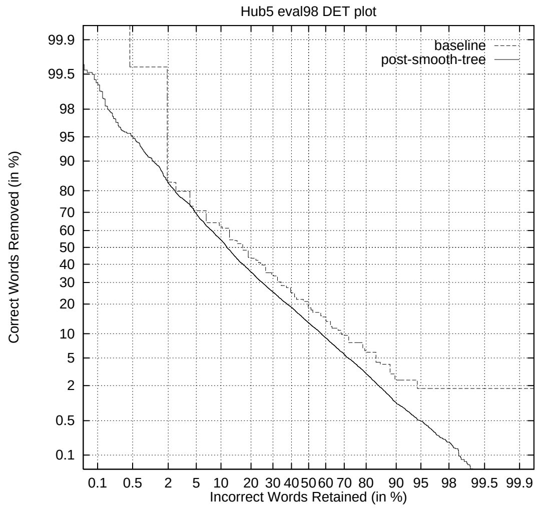

Figure 1: DET curves for Rover-combined systems (baseline and smoothed time-dependent posteriors)

6. CONCLUSIONS

In this paper we have discussed lattice based approaches to the estimation of word posterior probabilities and have presented a new method of incorporating these posterior distributions into a Viterbi decoder. We have also examined the application of word posteriors to the estimation of confidence scores.

The effectiveness of these techniques was demonstrated on the broadcast news and the conversational telephone speech corpora where improvements both in terms of word error rate and normalised cross entropy were achieved compared to the baseline HTK evaluation systems.

ACKNOWLEDGEMENTS

Gunnar Evermann is supported by scholarships from the EPSRC and the Cambridge European Trust.

7. REFERENCES

[1] G. Evermann. Minimum Word Error Rate Decoding. MPhil thesis, Cambridge University, 1999.

[2] P. Fetter, F. Dandurand, and P. Regel-Brietzmann. Word Graph Rescoring Using Confidence Measures. In *Proc. ICSLP'96*, pp. 10–13, Philadelphia.

[3] J.G. Fiscus. A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER). In *Proc. IEEE ASRU Workshop*, pp. 347–352, Santa Barbara, 1997.

[4] T. Hain, P.C. Woodland, T.R. Niesler, and E.W.D. Whittaker. The 1998 HTK System for Transcription of Conversational Telephone Speech. In *Proc. ICASSP'99*, pp. 57–60, Phoenix.

[5] L. Mangu, E. Brill, and A. Stolcke. Finding Consensus Among Words: Lattice-Based Word Error Minimization. In *Proc. Eurospeech'99*, pp. 495-498, Budapest.

[6] A. Stolcke, Y. König, and M. Weintraub. Explicit Word Error Minimization in N-Best List Rescoring. In *Proc. Eurospeech'97*, pp. 163–166, Rhodes.

[7] M. Weintraub. LVCSR Log-Likelihood Ratio Scoring for Keyword Spotting. In *Proc. ICASSP'95*, pp. 297–300, Detroit.

[8] F. Wessel, K. Macherey, and R. Schlüter. Using Word Probabilities as Confidence Measures. In *Proc. ICASSP'98*, pp. 225–228.

[9] F. Wessel, K. Macherey, and R. Schlüter. A Comparison of Word Graph and N-Best List Based Confidence Scores. In *Proc. Eurospeech'99*,pp. 315-318, Budapest.

[10] P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A. Tuerk, E.W.D. Whittaker, and S.J. Young. The 1997 HTK Broadcast News Transcription System. In *Proc. DARPA BN Transcription Workshop*, 1998.